Information Engineering

Topic 1:
Information organization and Thesauri

Anabel Fraga
Dept. of Informatics
Universidad Carlos III de Madrid

Leganés 2009

Information Engineering

What are we using for organization and retrieval?

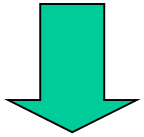
=> Language

The communication system used amongst humans is currently the

Language

Problem

- In a party, I take note in little papers everybody's telephone: If I keep the papers in a box…
  - How long will I need to recover a telephone?
    [I will need a lot of time to find it]
- If I keep them in an agenda...
  - How long will I need to recover a telephone?
    [Less Time]
  - But to write them in the agenda?
    [I will need more time to classify names and write them alphabetically]



- The time dedicated to classification minimizes time needed for retrieval. If I don't classify information I will need more time to recover it.
- Information → Classification → Information Retrieval

What is Classification?

- It is the distribution of a group of objects to various, generally distinct classes.
- Also is the coherent and structured system resulting from the previous operation.

Ordering and Classification

- Ordering and Classifying
  - Books
  - Photographs
  - Documents

Simple ordering (by shape, color, subject) is not sufficient for retrieval of information. Classification is also necessary

Classification Languages or Document Languages

- A classification language divides a real world domain to an ordered series of classes or sub-classes (e.g. Dewey Classification [330 for economics + 9 for geographic treatment + 4 for Europe = 330.94 European economy], Universal Decimal Classification [17:7 Relation of ethics to art])
- Document languages are designed for Library Science. It is a system of Controlled Vocabulary for representing document contents (e.g. thesauri or taxonomy) [e.g. Amazon].

Language Types

- Classification Languages, represent content in a synthetic form
- Indexing or combinatorial languages, permit the document content representation and the queries in analytic form
  - Controlled indexing language - Only approved terms can be used by the indexer to describe the document
  - Natural language indexing language - Any term from the document in question can be used to describe the document.
  - Free indexing language - Any term (not only from the document) can be used to describe the document.

Document Classification

- Mono-hierarchical
- Faceted
- Poly-hierarchical

Why controlled languages?

- Information classification starting from a list of consensual vocabulary

(by agreeing to a limited set of terms, users can benefit from a more efficient indexing and retrieval process)

What is indexing?

It is the activity of representing the content of a document or a query in analytic form, i.e. enumerating concepts and/or words.

IA-1 Index this document

Variants of Bagle in the Internet's 'Virus Soup'

EFE - SAN FRANCISCO (USA).- The appearance of **four new Bangle virus variants** has got security experts in "check mate". These worm variants threat by ending with ABC letters : **Bagle.Q** (which is the most spread), **Bagle.R, Bagle.S and Bagle.T**, smaller brothers of the original, have made their appearance in January.

The virus danger, that affects only systems that use the Windows operating system (and not the Apple Macintosh) comes from the fact that it does not require the user to open the file attached in the email in order to infect his computer, informs the EFE correspondent in USA Natalia Martín Cantero in an interview with Graham Cluley, a Sophos company consultant. Nevertheless, some experts believe that **the patches issued by Microsoft to combat this threat could be inefficient**, since computers could still be infected.

Hidden comments within programming code, have made experts conclude that computer pirates could be competing each other. Bagle.J contained a line within its code saying : "Hey, NetSky...don't ruin our business, wanna start a war? "

<u>Extract the terms you think that describe the document</u>

threat	January	business
appearance	virus danger	Netsky
Apple	expert	new
Bagle.J	security expert	original
Bagle.Q	file	patch
Bagle.R	Graham Cluley	computer pirate
Bagle.S	war	SAN FRANCISCO
Bagle.T	brother	system
code	smaller	Windows operating system
programming code	insufficient	Internet virus soup
comment	investigator	EFE
computer	ABC letter	user
EFE correspondent	line	Bagle variant
Sophos company consultant	Macintosh	gusano variant
email	Microsoft	Bagle virus variant
USA	Natalia Martín Cantero	

- Bagle Variant
- Internet Virus soup
- San Francisco
- Bagle virus variant
- Security experts
- Worm variants
- ABC letters
- Virus danger
- Windows operating system
- EFE correspondent
- Natalia Martín Cantero
- Grahan Cluley
- Sophos company consultant
- Programming code
- Computer pirate

IA-2 Manual Indexing

Internet Security
Virus
Worm
Bagle
Worm Variant
Windows O.S.
Computer pirate
Electronic mail
Netsky

Keep in mind:

- **Manual** indexing is achieved by **mental processing of concepts**, and **automatic indexing by the retrieval of words**.
- The **automatic** one usually is **more exhaustive** (lot of descriptors) but it **generates more noise**.
- The **consistency** between both is **low**

What is a Thesaurus?

- "It is the **vocabulary** of an indexing language, **controlled** and **formally organized**, aimed to do explicit the a priori **relationships** within **concepts**" (ISO 2788-1986)
  - Select descriptors of a domain
  - Establish relationships within descriptors

Elements of a Thesaurus

- A list of every important term representing a concept (single-word or multi-word) of the domain in interest. Preferred Terms - Descriptors
- A set of related terms for each term in the list. Not Preferred Terms (e.g. hounds USE dogs)
- Term Relationships
  - Equivalence (Related Terms RT)
    - a Synonym
    - a Quasi-Synonym
    - a Foreign Term
    - a Regional Term
    - an Archaic Term
  - Hierarchy (Broader Terms BT, Narrow Terms NT)
  - Association (MANSIO RT INN)
- *Scope Note (SN) (to avoid ambiguity for some terms. e.g. pool [the game] vs. pool [for swimming])*
- *Historical Data*
- *Facets (* next slide)*

Facets

Grouping of concepts of the same inherent category

Examples of categories that may be used for grouping concepts into facets are: **activities, disciplines, people, materials, living organisms, objects, places and times**. e.g.

- *animals*, *mice*, *daffodils* and *bacteria* could all be members of a **living organisms facet**;
- *digging*, *writing* and *cooking* could all be members of an **activities facet**;
- *Paris*, the *United Kingdom* and the *Alps* could all be members of a **places facet**.

Categories are normally chosen so that **facets are mutually exclusive**; a concept cannot then occur in more than one facet.

Facets may be subdivided into mutually exclusive **sub-facets**.

Facet Concepts

- Ranganathan was not the inventor of facet analysis, he is credited as the first to “systematize and formalize the theory”. It is said that Ranganthan's idea of a faceted classification scheme is inspired by a Lego-type toy set. Seeing that **the salesperson can build different toys just by combining the same pieces in a different way**, he builds his classification scheme by this analogy. (http://www.slais.ubc.ca/courses/libr517/winter2000/Group7/colon.htm)

Relationships in a Thesaurus

- Equivalence → For synonyms and near-synonyms
- Hierarchy → **Generic - Specific**, **Part - Whole**, **Enumerative**
- Association → Discipline, Instrumental, Cause, Attributive, Measure

Relationship Type	Relationship sub-type	Example
Hierarchy	Body organs	Skeleton - Articulations
	Geographic places	Andalucía-Cadiz
	Time	Week : (Monday, Tuesday.. etc)
	Social Structures	Armies – divisions – regiments
Synonyms	Jargon	Gun – piece
	Idiomatic Variations	Greece – Hellas
	Near-synonyms	Sea water – Salt water
Association	Discipline	Silviculture – Forests
	Instrument	Thermostat – temperature control
	Operation	Data processing – Automatic system
	Actions and its subject	Confinement – Prisoner
	Concept & property	Venoms – toxicity
	Origin	Rome – Romans
	Anti-agents	Plants – herbicides
	Units of Measurement	Electric current – Ampere

Thesauri types

- Thesauri can be :
- Monolingual 2788:1986[i]
- Multilingual ISO 5964: 1985[ii].

[i] Regulation UNE 50-106-90. Documentation Guidelines for the establishment and development of monolingual thesauri. Madrid. AENOR. 1990.

[ii] Regulation UNE 50-125-1997. Documentation Guidelines for the establishment and development of multilingual thesauri. Madrid. AENOR.1997.

Multilingual Thesauri

Multilingual thesauri can be formed using a **linguistic relationship**. In order to permit the implementation of other languages in the same thesaurus.

- The regulation defines a **source language** and a **series of target languages**.

General rules for the terms of a Thesaurus

- Nominal names or phrases (Noun Phrase, e.g. Red Sports Car) may be included
- Normally does not include proper names
- They should not be general and should not represent different thematic areas (theater of war – theater in arts)

Treating the terms

Directives	Examples
Plural for countable nouns	"TUBES"
Singular for materials	"WOOD"
Singular for processes, properties and conditions	"REFRIGERATION" "WEIGHT", "POVERTY"
Do not change the order	"RADAR ANTENNAS" (better than "ANTENNAS, RADAR")
Omit prepositions	"CARBOHYDRATE METABOLISM" (better than "METABOLISM OF CARBOHYDRATES")
Omit punctuation marks, abbreviations and special characters	"COOPERATIVE PROGRAMS" (better than "CO-OPERATIVE PROGRAMS" o "COOPERATIVE PROGRAMS") "MUSICAL NOTES" (better than "(MUSICAL) NOTES" o "MUS. NOTES"

Thesaurus Objectives

- Create a **map** of a **knowledge** area.
- Create a **standard vocabulary** for this area, assuring consistency in indexing and retrieval
- Ensure that **for every concept, only one term will be used** and none of its synonyms
- **Provide** the users with the ability to **find new concepts** through the system's relationships
- Serve as a **reference** for the users for the selection of a **correct term**
- The relationships between terms enhances retrieval

Thesaurus Example (Alphabetical Order)

Cava
SN: 01Cava from the Wines Concept
BT: Sparkly
RT: Champagne Glass
RT: Flaute Glass
RT: Pompadour Glass
Cava Glasses
BT: Wine Glasses
NT: Champagne Glass
NT: Flaute Glass
Chamomile
BT: Special Wines
RT: Fino
RT: Floral Aroma
RT: Generosos
Chamomile Aroma
BT: Floral Aroma

Thesaurus Example (Indexes)

Hierarchical Index

02 .Wines Classification
..Wines by Color
...Red
....Cubiertos
....Clarete
...White
...Roses
..Special Wines
...Noble
...Chamomile
...Fruity
...Mistella
...Sparkly
....Pickle
....Gasified
....Cava
...Must
....Yema
...Enverado
...Palo Cortado

Alphabetical Index

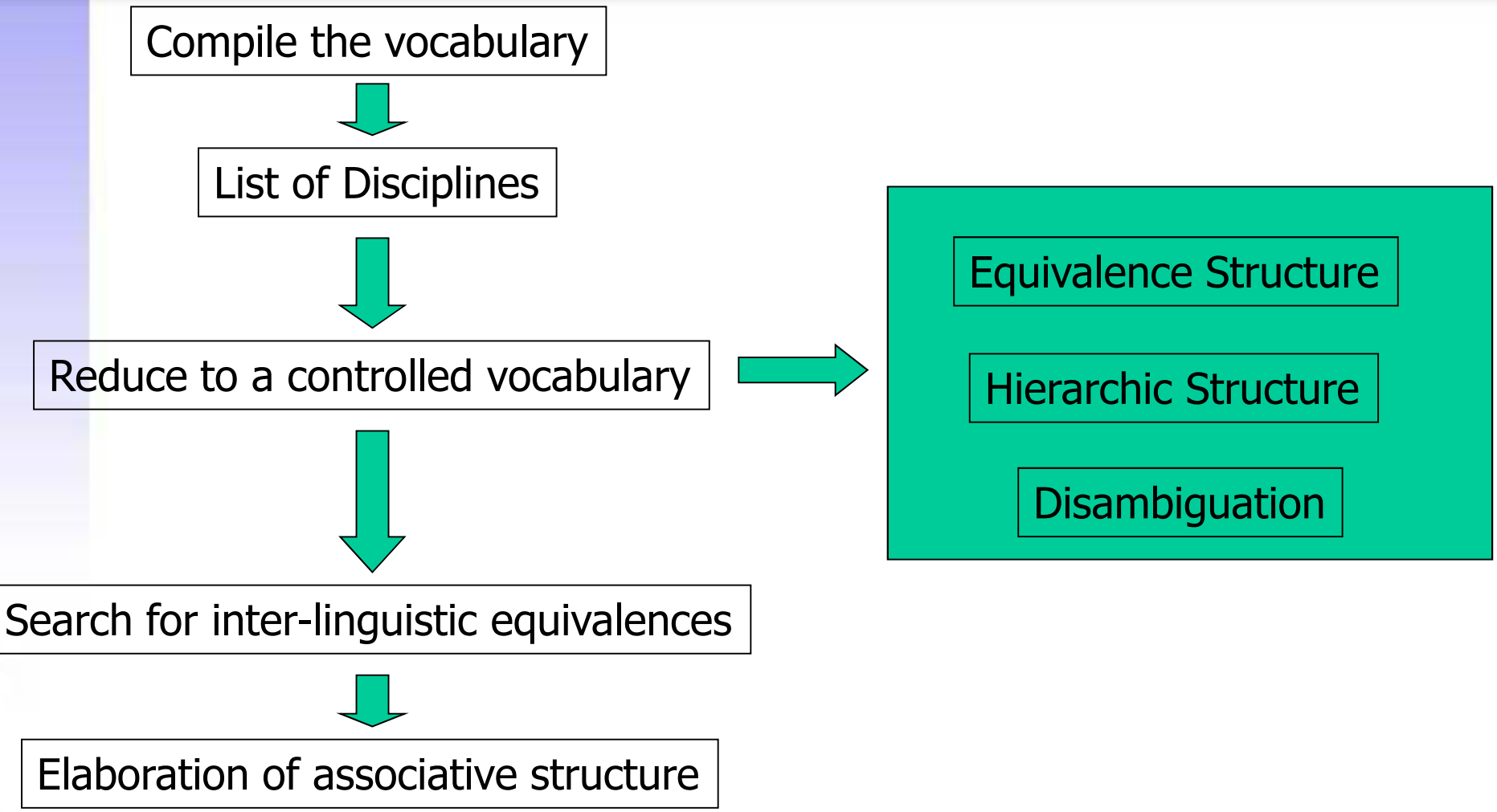
Cava
SN: 01Cava from the Wines Concept
BT: Sparkly
RT: Champagne Glass
RT: Flaute Glass
RT: Pompadour Glass

Cava Glasses
BT: Wine Glasses
NT: Champagne Glass
NT: Flaute Glass

Chamomile
BT: Special Wines
RT: Fino
RT: Floral Aroma
RT: Generosos

Chamomile Aroma
BT: Floral Aroma

How to construct a Thesaurus?



Benefits of Automatic Thesaurus Creation

- Greater Update
- Smaller cost and smaller construction time
- Knowledge and software reuse

Difficulties on Automatic Thesaurus Creation

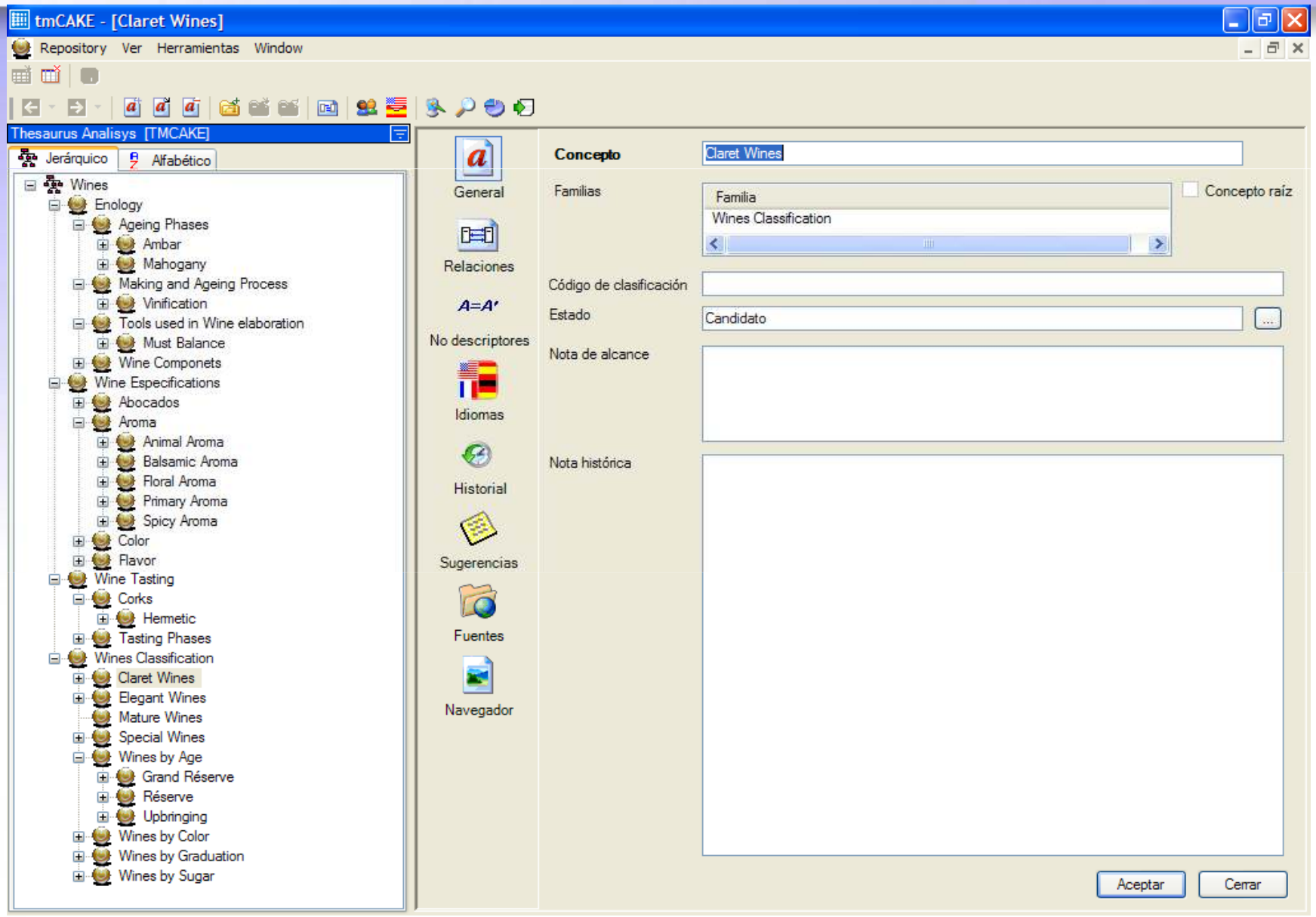
- Lack of Domain experts
- Indetermination on domain discrimination
- Automatic document treatment (preprocessing)
- Language Kind Type (formal or informal language)

Thesaurus Management Software

- A.k.a
- B.E.A.T
- CONCEPTO
- INDEX
- LIDOS
- PfleSaurus
- MIDOS
- MultiTes
- TAT
- Term Tree
- ThesMain
- THeW
- THSRS
- Noch mehr
- Thesauruss Software
- DomainREUSER

Thesaurus Management Software

Thesauri Generation Tool (tmCAKE / domainREUSER)



Thesauri Examples

- **Alcoholism**
  http://etoh.niaaa.nih.gov/dbtw-wpd/exec/dbtwpub.dll

- **MeSH**
  http://www.nlm.nih.gov/mesh/MBrowser.html

- **Art**
  http://www.getty.edu/research/tools/vocabulary/aat/index.html

- **cdu agency isbn**
  http://www.mcu.es/libro/plantilla?id=261&area=libro

- **Unesco Code mec**
  http://wwwn.mec.es/ciencia/jsp/plantilla.jsp?area=ayudaid&id=41

- **visual thesaurus**
  http://www.visualthesaurus.com/

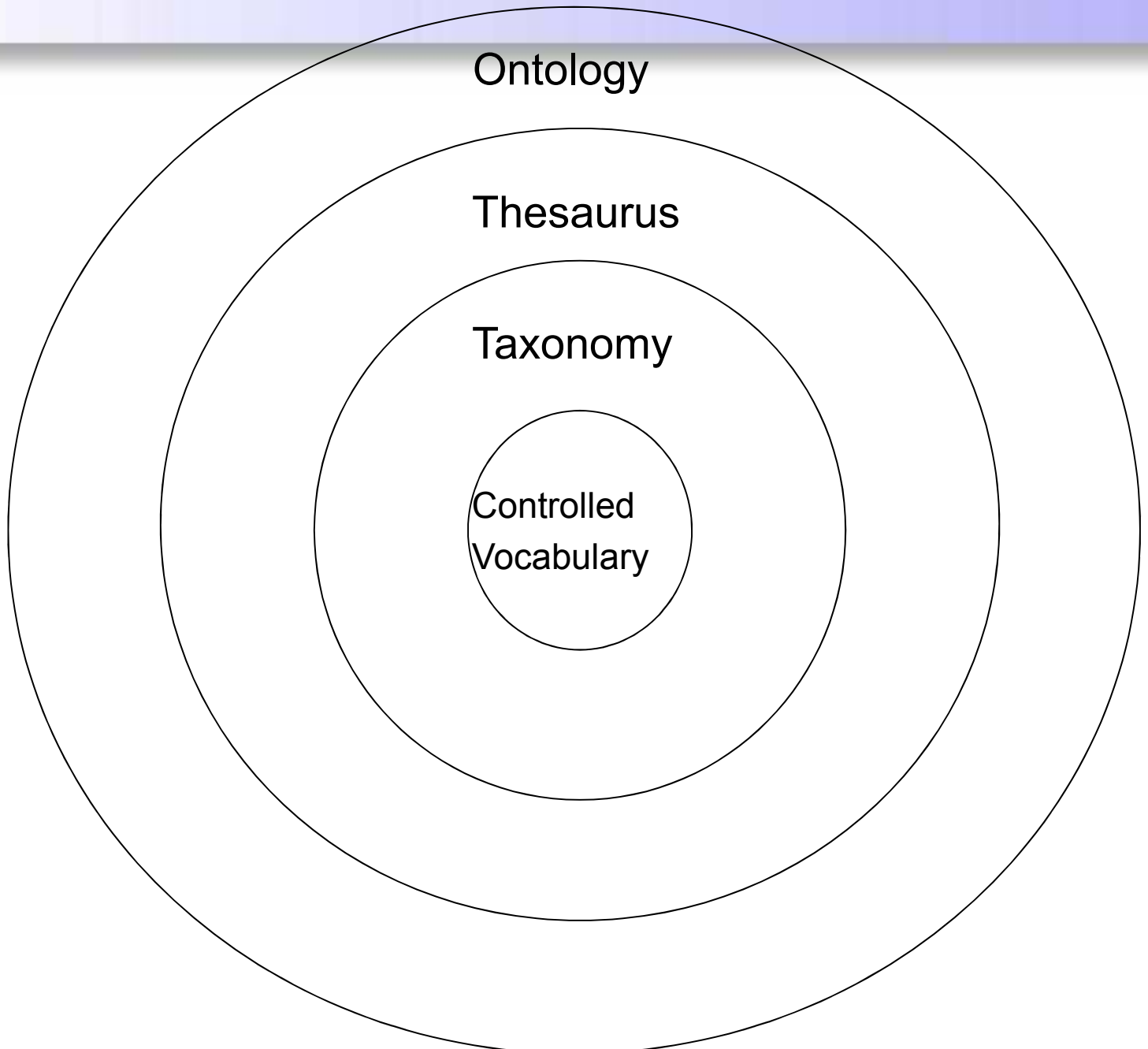
- **wordNet- Web MCR interface Meaning**
  http://nipadio.lsi.upc.edu/cgi-bin/wei4/public/wei.consult.perl

Conclusions

- Classification systems permit ordering and classification of elements and concepts in a physical and mental way
- Classification systems contribute to better information retrieval

A classification system may be used as a basic taxonomy for indexing new terms, so the new ones could be added to the old taxonomy creating a new enhanced one, in that case the reuse is amplified and knowledge will be enlarge little by little

Knowledge Structure



Bibliography

MSOffice9

- http://instruct.uwo.ca/677/thesaur/main04.htm (Tim Craven)
- Georges Van Slype. Indexing Languages. 1987
- ISO 2788: 1986
- NISO Z39:19 Structure and Organization of Information Retrieval Thesauri Standard. 1998
- Thesauri comparison http://willpower.demon.co.uk/thestabl.htm
- Thesauri Managers - Freeware http://publish.uwo.ca/~craven/freeware.htm
- Thesauri Managers http://www.fbi.fh-koeln.de/institut/labor/Bir/thesauri_new/thsoften.htm
- Cindoc Thesauri (Centro de Información y Documentación Científica) http://pci204.cindoc.es/tesauros/Index.html
- Glossary of terms relating to thesauri and other forms of structured vocabulary for information retrieval http://www.willpowerinfo.co.uk/glossary.htm

Diapositiva 36

MSOffice9 Los textos de la biblografía se debería traducir... creo...
; 27/10/2007

Exercise

- Manually construct a Wines thesaurus, by classifying relevant to the domain terms.
- Define the domain and the purpose of the Thesaurus.